

# Jednostavni linearni modeli

Regresija i analiza varijance (ANOVA)

Većina istraživanja nastoji opisati utjecaj neke nezavisne varijable (prediktora,  $x$ ) na zavisnu varijablu (ishod,  $y$ )

Npr:

*imaju li 2 hibrida kukuruza jednako dugačke klipove?*

*$x$  = hibrid*

*$y$  = duljina klipa*

*imaju li dulji gušteri veću masu od kraćih?*

*(utječe li duljina na masu guštera)*

*$x$  = duljina*

*$y$  = masa guštera*

Prediktori mogu biti kvantitativne (regresija) ili kategorijske varijable (ANOVA)

Generalni linearni model – poopćenje više tradicionalnih metoda (t – test, regresije, ANOVA, MANOVA, ANCOVA...)

Ishod(i) – kvantitativna varijabla (jedna ili više)

Prediktor(i) – kvantitativne ili kategorijske varijable (jedna ili više)

Y – matrica ishoda; X – matrica prediktora (dizajn matrica); B – matrica parametara (koeficijenata koje procjenjujemo); U- matrica pogrešaka (*noise*)

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}$$

Pretpostavke: greške su neovisne (*independent*) i multivarijatno normalno distribuirane

R funkcija:

***lm(y~x)***

Linearna regresija – opisuje (linearnu) ovisnost y o x varijablama

$$Y = \beta_0 + \beta_1 x_1 + \dots + \varepsilon$$

$\beta_0, \beta_1, \dots$  - koeficijenti (intercept + za svaki prediktor u modelu)

$\varepsilon$  - pogreška (reziduali – razlike izmjerenih i procijenjenih y vrijednosti)

Analiza varijance (ANOVA) – testira razliku (jednakost) prosjeka više grupa (prediktor je kategorijska varijabla)

The diagram illustrates the ANOVA model equation  $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ . Arrows point from descriptive text to each term in the equation:

- $y_{ij}$ : The individual's outcome
- $\mu$ : The average outcome over all individuals
- $\tau_i$ : The average outcome over all individuals in group i
- $\varepsilon_{ij}$ : The random noise that makes individual j unique

# Jedostavna linearna regresija – jedan ishod (y) i jedan prediktor (x)

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$\beta_0, \beta_1$  - koeficijenti (intercept + koeficijent za prediktor x)

$\varepsilon$  - pogreška (reziduali – razlike izmjerenih i procijenjenih y vrijednosti)

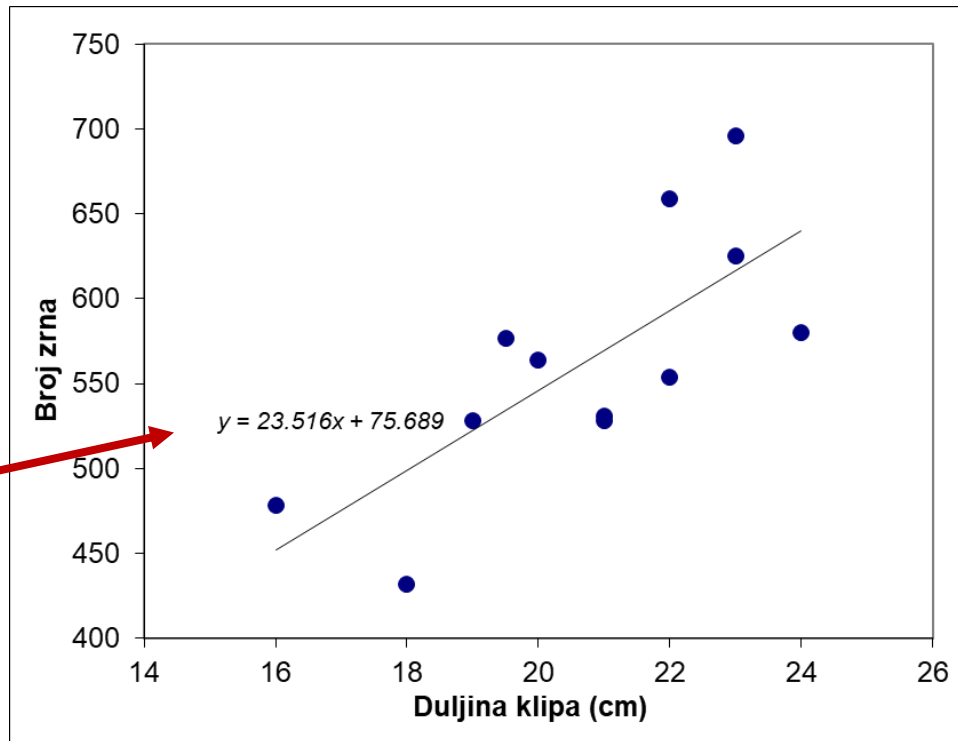
$$Y = a + bx \quad (\text{reg. pravac} - \text{„najbliži” točkama})$$

Procjena  $\beta_0 = a$  – intercept (odsječak na osi y)

Procjena  $\beta_1 = b$  – regresijski koeficijent (koeficijent smjera pravca)

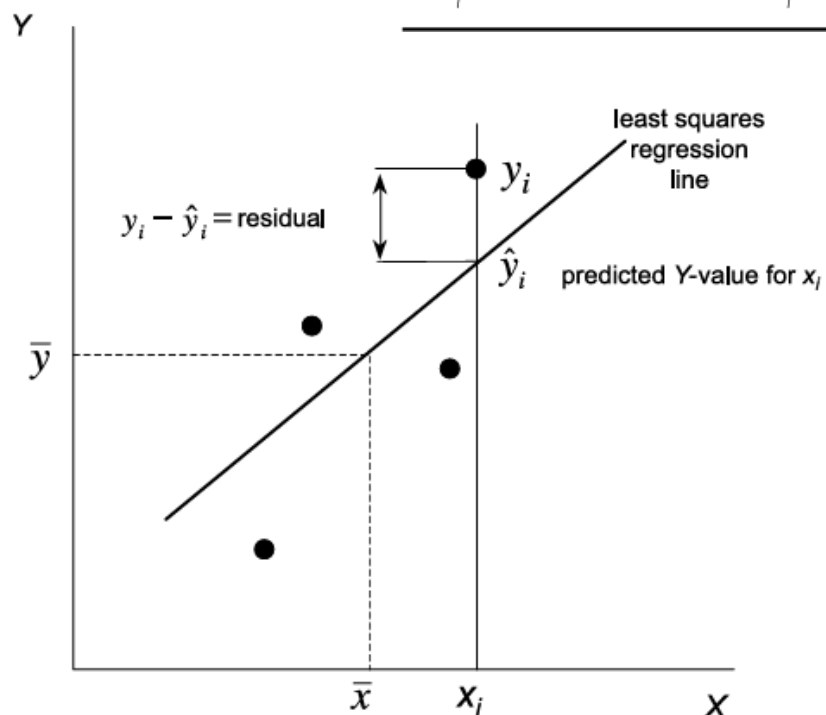
| Klip br. | Duljina klipa (cm) | Broj zrna |
|----------|--------------------|-----------|
| 1        | 21.0               | 528       |
| 2        | 21.0               | 531       |
| 3        | 24.0               | 580       |
| 4        | 23.0               | 625       |
| 5        | 22.0               | 554       |
| 6        | 16.0               | 478       |
| 7        | 19.5               | 577       |
| 8        | 22.0               | 659       |
| 9        | 23.0               | 696       |
| 10       | 19.0               | 528       |
| 11       | 18.0               | 432       |
| 12       | 20.0               | 564       |

Što znači 23.516?  
Što znači 75.689?



**Table 5.2** Parameters of the linear regression model and their OLS estimates with standard errors

| Parameter       | OLS estimate                                                                                 | Standard error                                                                                                        |
|-----------------|----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| $\beta_1$       | $b_1 = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^n (x_i - \bar{x})^2}$ | $s_{b_1} = \sqrt{\frac{MS_{\text{Residual}}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$                                        |
| $\beta_0$       | $b_0 = \bar{y} - b_1 \bar{x}$                                                                | $s_{b_0} = \sqrt{MS_{\text{Residual}} \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$ |
| $\varepsilon_i$ | $e_i = y_i - \hat{y}_i$                                                                      | $\sqrt{MS_{\text{Residual}}} \text{ (approx.)}$                                                                       |



Točka  $(\bar{x}, \bar{y})$  je sigurno na pravcu

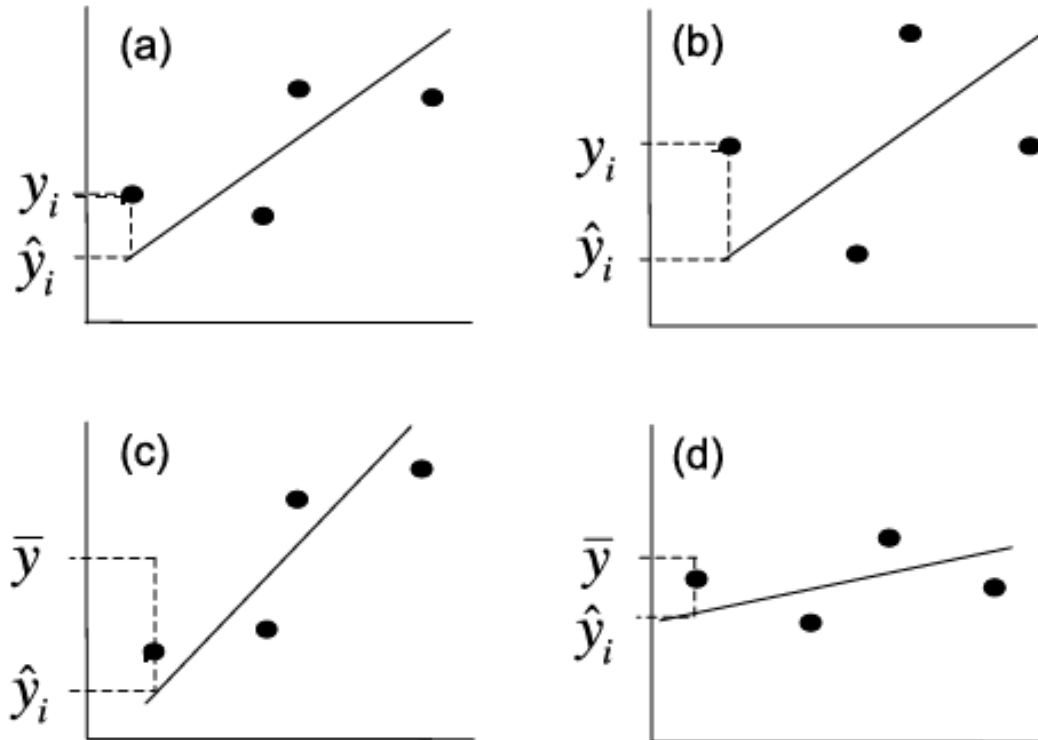
Na pravcu su procijenjene vrijednosti  $y$  (*kapa*)

Razlika izmjerenih i procijenjenih vrijednosti  $y$  – reziduali („pogreške“)

Reziduali su osnova za testove

**Table 5.3** Analysis of variance (ANOVA) table for simple linear regression of Y on X

| Source of variation | SS                                     | df      | MS                                               | Expected mean square                                              |
|---------------------|----------------------------------------|---------|--------------------------------------------------|-------------------------------------------------------------------|
| Regression          | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | 1       | $\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{1}$ | $\sigma_\varepsilon^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$ |
| Residual            | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$     | $n - 2$ | $\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$ | $\sigma_\varepsilon^2$                                            |
| Total               | $\sum_{i=1}^n (y_i - \bar{y})^2$       | $n - 1$ |                                                  |                                                                   |



**Figure 5.7** Illustration of explained and residual variation in regression analysis. Residual variation: (a) and (b) have identical regression lines but the differences between observed and predicted observations in (b) are greater than in (a) so the  $MS_{\text{Residual}}$  in (b) is greater than in (a). Explained variation: (c) and (d) have identical  $MS_{\text{Residual}}$  (the differences between the observed and predicted values are the same) but the total variation in Y is greater in (c) than (d) and the differences between the predicted values and the mean of Y are greater in (c) than in (d) so  $MS_{\text{Regression}}$  would be greater in (c) than in (d).

| Table 5.3   Analysis of variance (ANOVA) table for simple linear regression of Y on X |                                        |         |                                                  |                                                                   |
|---------------------------------------------------------------------------------------|----------------------------------------|---------|--------------------------------------------------|-------------------------------------------------------------------|
| Source of variation                                                                   | SS                                     | df      | MS                                               | Expected mean square                                              |
| Regression                                                                            | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | 1       | $\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{1}$ | $\sigma_\varepsilon^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$ |
| Residual                                                                              | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$     | $n - 2$ | $\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$ | $\sigma_\varepsilon^2$                                            |
| Total                                                                                 | $\sum_{i=1}^n (y_i - \bar{y})^2$       | $n - 1$ |                                                  |                                                                   |

Total – ukupna varijabilnost varijable y – može se „rastaviti” na model+reziduale

Model (regresija) – varijabilnost varijable y objašnjena regresijom (odnosom s x) – objašnjava koliko dobro model opisuje/predviđa  $\bar{y}$ )

Reziduali (pogreška) – dio varijabilnosti y varijable koji nije opisan regresijom – razlika procijenjenih i izmjerenih vrijednosti  $y_i$  – mjerilo udaljenosti izmjerenih vrijednosti y od reg. Pravca

Koeficijent determinacije –  $R^2$  – udio ukupne varijance y varijable objašnjen regresijom

$$R^2 = \text{SS}_{\text{regresije}} / \text{SS}_{\text{total}} = 1 - (\text{SS}_{\text{reziduala}} / \text{SS}_{\text{total}})$$

## Testovi značajnosti

Testira se nulta hipoteza:  $H_0: \beta_1=0$

(y se ne mijenja ako se mijenja x)

### 1) F-test (omjer varijanci)

$$F = MS_{Model} / MS_{Rezidual}$$

F se usporedi sa F distribucijom (za  $1; n-2$  stupnjeva slobode) – dobije se  $p$  – vjerojatnost prihvatanja  $H_0$

### 2) t- test (omjer procjenitelja i njegove pogreške)

$$t = b / s_b$$

t se usporedi sa t distribucijom (za  $n-2$  stupnjeva slobode) – dobije se  $p$  – vjerojatnost prihvatanja  $H_0$

Isti zaključak –  $F=t^2$

## Pretpostavke (pojednostavljeno)

- 1) Normalnost – Pretpostavlja se da su reziduali normalno raspodijeljeni (distribuirani)
- 2) Neovisnost – Pretpostavlja se da su vrijednosti Y varijable i pogreške neovisne jedne od drugih (nema autokorelacije) – često je narušeno kod vremenskih serija ili prostornih podataka
- 3) Homogenost varijanci – varijance Y varijable i pogreške su homogene za svaku razinu x varijable – bitnije kod modela ANOVA-e (kod regresije uglavnom nemamo više Y varijanata za svaku razinu x varijable)

Ove pretpostavke se najjednostavnije provjeravaju (ne egzaktno) dijagnostičkim grafovima (vidi primjere)

U slučaju narušenih pretpostavki se mogu koristiti nelinearni (ili neki drugi) modeli ili se varijable mogu transformirati (preračunati – npr logaritmirati), no pri tome treba biti jako oprezan (što često nije slučaj)

Primjeri

<https://antoinesoetewey.shinyapps.io/statistics-202/>

<https://shiney.zoology.ubc.ca/whitlock/Residuals/>

<https://mjkeough.github.io/examples.html>

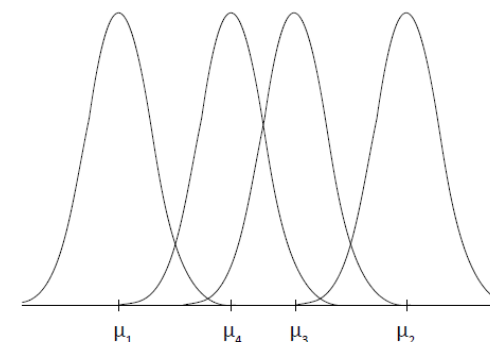
Analiza varijance (ANOVA) – Rastavljanje varijance na komponente (već smo koristili kod regresije)

Linearni model sa kategorijskim prediktorom (x) – (*as.factor()* u R-u)

Zapravo uspoređuje prosjeke varijanata y varijable grupirane prema x varijabli

*Nulta hipoteza:*

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_i$$



Primjer – jesu li duljine klipova kukuruza (y) jednake za 3 hibrida (x)

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

The individual's outcome

The average outcome over all individuals

The average outcome over all individuals in group i

The random noise that makes individual j unique

Primjer – izmišljeni podatci za duljinu klipa kukuruza 3 hibrida (zaokruženo na cm)

| Podatci ( $y_{ij}$ ) |    |    | Procjene prosjeka<br>populacija tretiranja ( $\bar{y}_{i.}$ ) |     |     | Odstupanja podataka<br>od procjena ( $\varepsilon_{ij}$ ) |    |    |
|----------------------|----|----|---------------------------------------------------------------|-----|-----|-----------------------------------------------------------|----|----|
| A                    | B  | C  | A                                                             | B   | C   | A                                                         | B  | C  |
| 18                   | 17 | 10 | 15                                                            | 18  | 9   | 3                                                         | -1 | 1  |
| 12                   | 19 | 6  |                                                               |     |     | -3                                                        | 1  | -3 |
| 16                   | 15 | 7  |                                                               |     |     | 1                                                         | -3 | -2 |
| 13                   | 18 | 11 | ...                                                           | ... | ... | -2                                                        | 0  | 2  |
| 14                   | 16 | 12 |                                                               |     |     | -1                                                        | -2 | 3  |
| 17                   | 21 | 8  |                                                               |     |     | 2                                                         | 3  | -1 |
| 15                   | 20 | 9  | 15                                                            | 18  | 9   | 0                                                         | 2  | 0  |

| Izvori<br>varijabilnosti | $df$       | $SS$                            | $MS$                        | $F$                          |
|--------------------------|------------|---------------------------------|-----------------------------|------------------------------|
| Ukupno                   | $tr - 1$   | $SS_{UKUPNO} = SS_{RM}$         |                             |                              |
| Tretiranja               | $t - 1$    | $SS_{TRET} = SS_{RM} - SS_{FM}$ | $\frac{SS_{TRET}}{t - 1}$   | $\frac{MS_{TRET}}{MS_{ERR}}$ |
| Pogreške<br>(reziduali)  | $t(r - 1)$ | $SS_{ERR} = SS_{FM}$            | $\frac{SS_{ERR}}{t(r - 1)}$ |                              |

| Izvori<br>varijabilnosti | $df$ | $SS$ | $MS$ | $F$   | $F_{tab}$<br>$p = 5\% \quad p = 1\%$ |      |
|--------------------------|------|------|------|-------|--------------------------------------|------|
| Ukupno                   | 20   | 378  |      |       |                                      |      |
| Tretiranja               | 2    | 294  | 147  | 31,50 | 3,55                                 | 6,01 |
| Pogreške<br>(reziduali)  | 18   | 84   | 4,67 |       |                                      |      |

$$SS_{UKUPNO} = SS_{TRET} + SS_{ERR}$$

Što dokazujemo ne prihvatanjem nulte hipoteze?

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_j$$

Ukoliko nas zanima koje grupe (tretmani, razine x varijable) se međusobno razlikuju treba provesti dodatne testove (usporedbe parova, pairwise comparisons)

Definirane unaprijed – prije analize – kontrasti

Definirane tek nakon odbacivanja  $H_0$  – post hoc usporedbe – t-testovi –  $t = D/s_D$

SS reziduala je sastavljena od pogrešaka svih grupa – koristimo ju za procjenu  $s_D$

Usporedbe sa kontrolom, ili svih razina sa svim razinama

Potrebno je korigirati  $p$  vrijednosti takvih testova radi višestrukih usporedbi (Bonferronijeva korekcija, False discovery rate...)

## Density-dependent decline of early horn growth in European mouflon

Krešimir Kavčič<sup>a</sup>, Luca Corlatti<sup>b,\*</sup>, Toni Safner<sup>c,d</sup>, Ivan Gligora<sup>e</sup>, Nikica Šprem<sup>a</sup>

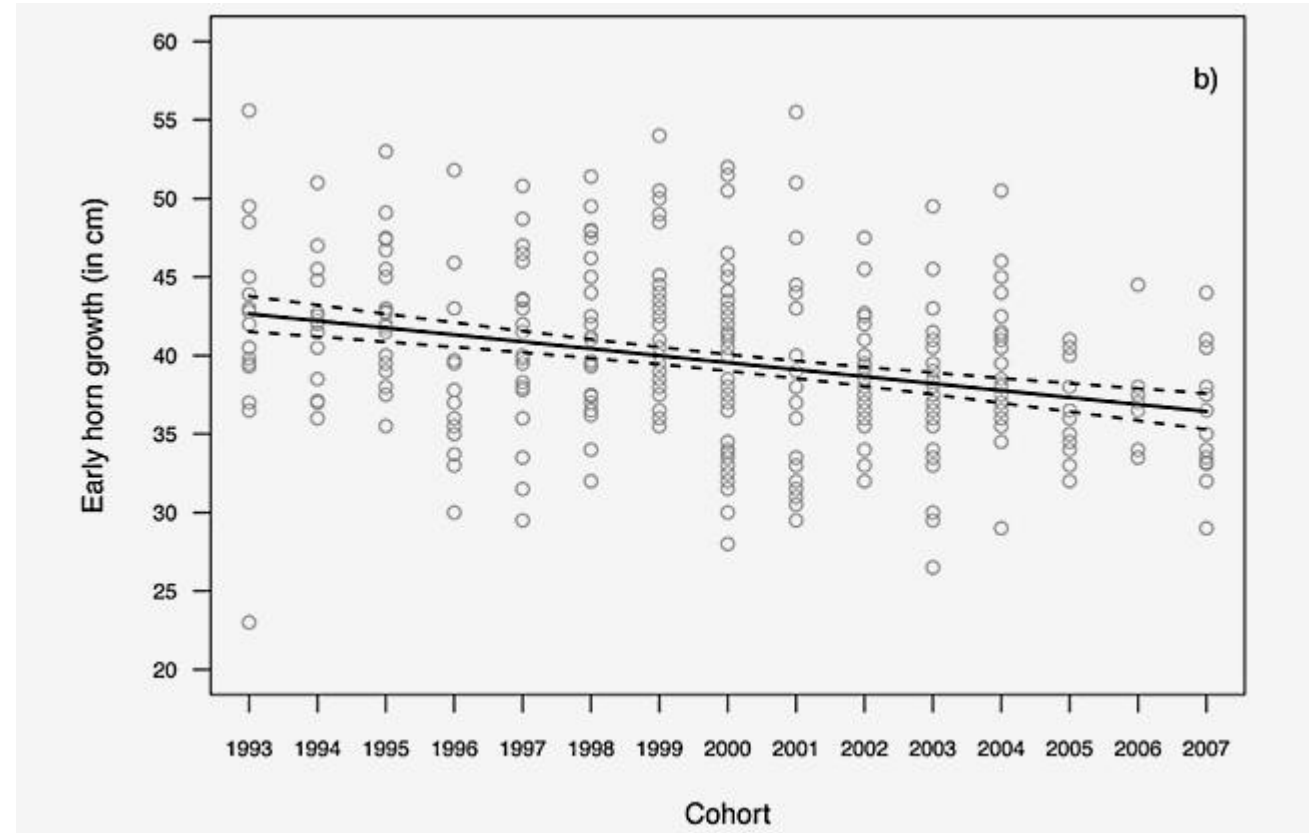
Rani rast roga muflona

n=341

y=rani rast roga (rast u drugoj i trećoj godini)

x=kohorta (godina okota)

Ako je x kategorička ali numerička varijabla (godine, doze i slično) koja je razlika u zaključcima između ANOVA-e i regresijske analize??



Primjeri

<https://istats.shinyapps.io/ANOVA/>

<https://mjkeough.github.io/examples/medley.nb.html>

I u regresiju i u ANOVA-u možemo „dodavati” prediktore

Multipla regresija

Faktorska ANOVA (interakcije!!)

ANCOVA (kvantitativni i kategorijski prediktor - interakcije)

„Slaganje” efekata – crossed ili nested efekti

Fiksni efekti – razine faktora su unaprijed definirane, testira se razlika upravo između tih razina (sva „ciljana” istraživanja)

Slučajni efekti – razine su izvan kontrole (godine i sl.)